Формирование многоязычных словарей и их использование при кросс-языковом поиске информации

Н.Н. Абрамова Е.И. Глобус

Федеральное государственное унитарное предприятие «Научно-исследовательский центр информатики при МИД Российской Федерации»

NAbramova@mid.ru

Аннотация

Рассматриваются проблемы поиска иноязычной информации по запросам на русском языке, так называемый кросс-языковой поиск. Ранее подобные исследования проводились при создании систем машинного перевода только для автоматического составления двуязычных словарей группой исследователей под руководством профессора Г.Г. Белоногова. Суть исследования состоит в разработке методов автоматического составления многоязычных словарей по заголовкам параллельных текстов и методов перевода запросов с русского языка на несколько иностранных языков с помощью многоязычных словарей. В результате исследования сформирован машинный многоязычный словарь объемом около 20.000 словарных единиц, который получен на основе автоматически составленного словаря и дополнения из традиционных словарей. Проведен эксперимент по переводу реальных пользовательских запросов к поисковой системе Яндекс, который показал эффективность разработанных методов.

Currently, the issues of multilingual information retrieval through textual inquiries in Russian, or the so-called cross-language information retrieval, are being considered. Earlier, a team of researches under direction of professor G.G.Belonogov conducted similar researchers aimed solely at automatic compiling of bilingual dictionaries when building up machine translation systems. The research focuses on developing methods of automatic compilation of multilingual dictionaries according to the headings of parallel texts as well as methods of translation of inquiries from Russian into several foreign languages using multilingual dictionaries. The research resulted in the electronic multilingual dictionary compiled about 20.000 entries that had been produced on the basis of automatically compiled dictionary and supplements from traditional dictionaries. An experiment in translation of real user inquiries to Yandex retrieval system showed the effectiveness of the developed methods.

1. Введение (постановка задачи, обзор существующих методов решения)

В современном мире, охваченном глобализацией и интеграцией стран и народов, возрастает роль многоязычных словарей. К примеру, расширение ЕС - это не только политическая и экономическая проблемы, но и лингвистическая проблема. Сегодня в ЕС говорят на 20 официальных языках, относящимся к разным языковым семьям: германским, романским, славянским, угро-финским. В рамках ЕС ведется разработка многоязычных словарей, которые будут помогать в общении. Однако роль многоязычных словарей не исчерпывается функцией перевода. Они применяются для обучения языкам, унификации терминологии в различных областях науки и технологий, поиска информации в иноязычных базах данных по запросам на родном языке (так называемый кросс-языковой поиск). В последние годы в связи с бурным развитием интернет-технологий актуальность многоязычной лексикографии резко возрастает.

Хотя в России достаточно распространена точка зрения, что все грамотные пользователи Интернета должны понимать английский язык как язык межнационального общения, и этого вполне достаточно, чтобы найти любую информацию, на самом деле все обстоит иначе. Приведем данные из статьи американского аналитика Роберта Левина [11]. Свыше 800 миллионов, т.е. более 65 % всех интернет-пользователей не говорят по-английски. Причем, ежегодный прирост не говорящих по-английски интернет-пользователей составляет свыше 140 миллионов человек. По крайней мере половина всех текстов в Интернете написана не на английском языке. Эти факты красноречиво свидетельствуют об актуальности создания поисковых систем с кросс-языковыми возможностями.

Такие системы пока не получили в Интернете распространения. Компания ПРОМТ выпустила первую систему машинного перевода для Интернета под названием WebTranSite, снабженную средством перевода запросов на поиск с английского, немецкого и французского языков на русский и обратно. Однако качество перевода не удовлетворяет взыскательных пользователей. То же самое можно сказать и о системе перевода SYSTRAN (Systran, USA) на сервере Altavista, которая, правда, обладает большими возможностями в смысле количества задействованных языков.

Для поиска в базах данных в мире популярна разработка компании Convera Technologies - программный продукт

RetrievalWare (RW), в котором реализована функция кросс-языкового поиска. Пользователь может формировать запросы на одном из 14 языков (русском, английском, немецком, французском, испанском, арабском, китайском, японском и т.д.), а получать в ответ документы на других языках. Однако сами разработчики утверждают, что использовать эту дорогостоящую систему только для работы с Internet-ресурсами нецелесообразно. Поэтому основные клиенты Convera - крупные банки, государственные организации, СМИ, исследовательские центры и т.д.

Из наших отечественных разработок следует отметить систему поиска информации в русскоязычных реферативных базах данных по запросам на английском языке, которая уже несколько лет функционирует в ВИНИТИ РАН [3]. Для перевода запросов с английского языка на русский используется система ERTRANS, затем русскоязычные переводы формализуются с помощью логических операторов и исключается малоинформативная лексика для использования их в ИПС «Сокол».

Основную часть трудозатрат при разработке кросс-языкового поиска составляют затраты на формирование словарей. Электронные словари отличаются от традиционных книжных словарей прежде всего своим назначением. Они предназначены для использования в автоматизированном режиме для поиска информации в электронных базах данных, массивах документов и т.д. Вследствие этого лексика словарей должна в максимальной степени быть приближена к лексическому составу текстов, в которых проводится поиск. Этой цели можно достигнуть, если составлять словари по реальным текстам. Составление словарей – сложный и трудоемкий процесс, а если речь идет о многоязычных словарях, то трудности возрастают многократно.

При разработке многоязычных словарей требуется привлечение коллективов разработчиков, состоящих из профессионалов, владеющих несколькими языками.

Существующая в настоящее время практика разработки многоязычных машинных словарей опирается на традиционные книжные словари и тезаурусы, которые сначала с помощью технологий сканирования и распознавания текстов переводятся в электронную форму, а затем автоматически записываются в требующемом формате. По такому методу созданы все известные словари, например Lingvo, MultiLex, Eckado и т.д. Некоторые многоязычные словари (например, во ВНИИКИ Госстандарта) создавались вручную на основе справочников, пособий, толковых

словарей, энциклопедий и т.д., а затем записывались в электронной форме.

Однако, если использовать только лексику, содержащуюся в традиционных книжных словарях, то получить хорошее покрытие текстов невозможно, так как словари всегда отстают от реальных текстов.

Оригинальный способ решения этой проблемы предлагался компанией Multitran в проекте multitran.ru - ведение словарей с помощью Интернет в режиме on-line [8].

Мы видим путь решения в том, чтобы при разработке машинных многоязычных словарей наряду с обработкой существующих традиционных словарей составлять словари по реальным текстам, естественно, используя автоматизированные методы.

Цель настоящего исследования состоит в разработке методов и алгоритмов автоматизированного составления многоязычных словарей и использования их при поиске информации в иноязычных информационных ресурсах по запросам на русском языке.

Идея составления двуязычных словарей на основе билингв (или параллельных текстов) впервые была высказана профессором Белоноговым Г.Г. и реализована его учениками. В настоящее время в компьютерной лингвистике активно развивается направление по исследованию многоязычных корпусов текстов [4].

В 90-х годах прошлого века в ВИНИТИ был составлен русско-английский словарь на основе массива текстов заголовков научно-технических документов на английском языке и их переводов на русский язык. Словарь составлялся с использованием средств автоматизации во всех основных операциях, таких как выделение пар русских и английских заголовков и их нумерация, морфологический анализ русских словосочетаний, построение поисковых образов словосочетаний[1] (ПОС) и сортировка словосочетаний по ПОС-ам, удаление дублирующихся пар заголовков. Переводные соответствия между русскими и английскими словосочетаниями устанавливались с помощью дистрибутивно-статистического метода. В основе этого метода лежит гипотеза о том, что *«если два предложения (на русском и английских языках) являются переводами друг друга, то для каждого слова и/или словосочетания одного из предложений с высокой вероятностью найдется эквивалентное ему по смыслу словосочетание или слово из другого предложения, и при этом переводы слов, входящих в состав русского словосочетания, будут располагаться в английских предложениях компактно»* [3].

Сначала по всему массиву русских предложений формируется частотный словарь словосочетаний с сохранением номеров предложений, содержащих эти словосочетания. Затем для каждого словосочетания, удовлетворяющего некоторому порогу частотности, формируется массив соответствующих английских предложений, по которым составляются частотные словари английских словосочетаний. Одно или два самых частых словосочетаний выбираются в качестве перевода искомого русского словосочетания. Этот алгоритм был апробирован на 1 млн. пар англо-русских заголовков и показал вполне приемлемые результаты.

Попыток автоматического составления многоязычных словарей в литературе не описано.

2. Идея исследования

Основная идея исследования состоит в том, чтобы разработать дистрибутивно-статистический метод для составления многоязычных словарей в автоматизированном режиме. Если сравнивать составление двуязычных и многоязычных словарей, то для последних много неясного. Ранее шла речь о гипотезе Белоногова Г.Г., которая была положена в основу составления двуязычных словарей по заголовкам билингв. Предстоит проверить эту гипотезу для текстов на четырех языках: русском, английском, немецком и французском.

Здесь более остро стоит вопрос о многозначности слов и словосочетаний. Известно, что в двуязычных словарях в силу объективных языковых факторов (существование синонимии слов и словосочетаний, грамматических конструкций, синонимичных аббревиатур для выражения одного и того же понятия) не всегда достигается взаимно-однозначное соответствие русских и иноязычных словосочетаний. В многоязычных словарях роль этих факторов усиливается за счет включения новых языков, в каждом из которых имеются эти явления. Поэтому при составлении словарей нужно либо изначально выявлять синонимию в исходных параллельных текстах с помощью многоязычного тезауруса, приводя все синонимы к заглавному дескриптору, либо выявлять многозначность слов и словосочетаний в процессе анализа текстов.

Предстоит проверить вторую часть гипотезы о контактном расположении слов в переводных эквивалентах на разных языках. Если это положение не всегда выполняется, то нужно найти способы выбора переводных эквивалентов.

Другой аспект проблемы заключается в отыскании исходного материала для разработки словарей. Очевидно, что, как и сам

многоязычный словарь, так и исходные параллельные тексты, не должны быть ориентированы на какие-то определенные тематические области в силу политематичности сети Интернет.

Пользователи Рунета чаще всего в иноязычных ресурсах ищут информацию по политике, бизнесу, праву, программному обеспечению, культуре, спорту и т.д. [9].

Источником параллельных текстов также является Интернет. В настоящее время в сети есть много зеркальных сайтов с различной тематической направленностью. Приведем в качестве примера сайты, информация которых может быть успешно использована при составлении многоязычных словарей: сайт МИД России (русский, английский, немецкий, французский, испанский, китайский языки), сайт МГУ (русский, английский, немецкий, французский), сайт Европейского Союза (20 языков членов ЕС), сайт компании Самсунг (русский, английский, немецкий, французский, испанский, корейский, китайский и т.д.).

Что касается второй части исследования - перевода пользовательских запросов с помощью многоязычного словаря, то здесь необходимо провести анализ и формализацию пользовательских запросов. В основе формализации лежат принципы синтаксического анализа запросов. Для вычленения из текста запросов наименований понятий можно использовать те же методы, что и при составлении словарей. Для нахождения переводных соответствий выделенным из русскоязычных запросов словосочетаниям нужно провести их поиск в многоязычном словаре. Найденные иноязычные переводы необходимо записать в формате, который годился бы для большинства поисковых систем.

3. Описание методов, алгоритмов и экспериментов

3.1. Метод автоматического формирования многоязычных словарей по параллельным текстам заголовков

Задача состоит в выделении из русского предложения смыслов (наименований понятий), выраженных словами и словосочетаниями. То есть для каждого слова и словосочетания из предложения на русском языке необходимо получить его дистрибуцию - перечень слов и словосочетаний на остальных трех языках, встречающихся в переводах предложений, в которые входит искомое слово или словосочетание. Очевидно, что если бы для каждого русского словосочетания существовало единственное его выражение на каждом из иностранных языков, то частота появления соответствующего ему переводного эквивалента в

точности была бы равна количеству русских предложений, в которых встречается данное словосочетание. Как правило, в реальных текстах это условие не выполняется и приходится устанавливать частотный порог опытным путем, например, равным 1/2 или 2/3 от количества предложений. Таким образом, отбор переводных эквивалентов для русского словосочетания должен проводиться среди иноязычных словосочетаний, частота появления которых в его дистрибуции не менее выбранного порога. Словосочетание с самой высокой частотой появления и будет искомым переводным эквивалентом.

Самым сложным вопросом является определение границ словосочетаний, как в русских так и в английских предложениях. Для русского языка нами использовался метод, базирующийся на морфологическом и синтаксическом анализе текстов [1]. Здесь мы не будем рассматривать вопросы построения алгоритма морфологического анализа[2]. Дадим лишь краткое описание алгоритма выделения словосочетаний из русских текстов.

Так как исходные тексты являются заголовками документов, то речь идет об именных словосочетаниях, которые представляют собой цепочки связанных по смыслу и контактно расположенных слов, относящихся к грамматическим классам «существительные», «прилагательные», «предлоги», «наречия» и «сочинительные союзы». Предлоги и союзы не могут стоять в начале и в конце словосочетания.

Границы между словосочетаниями проводились по знакам препинания (исключая запятую между однородными членами и точку после инициалов и сокращений), по любым частям речи, отличным от существительных, прилагательных, наречий, предлогов и союзов, а также по существительным или прилагательным в именительном или винительном падеже без предшествующего предлога.

Имена и фамилии, географические названия, названия организаций, партий распознавались в текстах с помощью специальных словарей и считались как отдельные словосочетания, за исключением фамилий, входивших в состав терминов (например, МГТУ имени Н.Э. Баумана, ряд Тейлора, цепи Маркова, преобразование Фурье, поправка Джексона-Вэника и т.д.). Для таких названий имеется отдельный словарь.

Все словари собственных названий имеют одинаковую структуру, состоящую из собственно названия и его поискового образа, который необходим для отождествления словосочетаний из текста и словаря. Поисковый образ словосочетания формируется

автоматически по результатам морфологического анализа и состоит из словоизменительных основ входящих в него слов. Словосочетания считаются тождественными по смыслу, если ПОС-ы у них полностью совпадают. Например, у словосочетаний *автоматическая обработка данных, автоматической обработки данных, автоматическую обработку данных* один ПОС *автоматическ обработк данн.* При таком построении ПОС-ов словосочетания, отличающиеся порядком следования слов или синонимичной синтаксической конструкцией (при переходе слов из одних синтаксических классов в другие), не будут считаться тождественными. Однако такие языковые факторы в именах собственных практически не встречаются. Каждому словосочетанию ставился в соответствие номер предложения (все предложения-переводы имеют тот же номер), в котором оно встретилось. Затем по всему массиву выделенных словосочетаний с приписанными номерами предложений строился инверсный файл, в котором словосочетание сопровождалось всеми номерами предложений, в которые оно было включено.

Для выделения словосочетаний из английских, немецких и французских текстов можно было бы разработать аналогичные алгоритмы, если бы мы располагали процедурами морфологического анализа для всех языков словаря. Вследствие отсутствия таковых, мы использовали отсечение окончаний слов и дистрибутивно-статистический метод для установления близости между словами.

Для всех указанных языков были сформированы массивы границ словосочетаний, куда вошли знаки препинания (кроме точки и запятой), фамилии, географические названия и другие имена собственные и массивы союзов и предлогов, которые не могут стоять в начале и в конце словосочетаний. Выделение словосочетаний из иноязычных текстов проводилось в следующем порядке:

1. Для каждого русского словосочетания выбирались все тройки иноязычных предложений, номера которых совпадали с номерами включающих его русских предложений.
2. Вначале в каждом из выбранных предложений выявлялись слова с одинаковыми словоизменительными основами с помощью грамматических таблиц и отсекались окончания (например, agression-s (анг.), Aggression-nen (нем.), aggression-s (фр.)).

Затем по всем этим предложениям составлялись частотные словари английских, немецких и французских слов.

3. В частотных словарях отбирались слова с частотой, не меньшей 2/3 или ½ (в зависимости от языка) от количества предложений.
4. Согласно гипотезе о контактности расположения слов в переводных эквивалентах, выбирались непрерывные последовательности (цепочки), состоящие из отобранных частых слов с длиной не менее двух слов.
5. Проводилось разбиение цепочек, если какие-то слова находились в словаре границ словосочетаний.
6. Исключались предлоги и союзы, стоящие в начале и в конце цепочек. Отредактированные цепочки слов рассматривались в качестве словосочетаний.

На следующем этапе составлялись частотные словари словосочетаний на английском, немецком и французском языках, относящиеся к указанной выборке предложений (п.1).

В качестве иноязычного переводного эквивалента выбиралось словосочетание с самой высокой частотой, а при наличии нескольких таких словосочетаний выбирались все. Наличие нескольких переводных соответствий говорит либо о синонимии словосочетаний, либо о погрешности метода.

3.2. Эксперимент по автоматическому составлению многоязычного (русско-англо-немецко-французского словаря)

Для автоматического составления словаря были использованы тексты пресс-конференций, заявлений МИД России, сообщений для печати, посланий по внешнеполитическим вопросам, выступлений Президента, Председателя Правительства, выступлений руководителей МИД, выступлений руководителей российских делегаций на международных форумах (начиная с 2001 года по июль 2005 г.). Все эти материалы находятся в открытом доступе на официальном сайте МИД России (www.mid.ru). Из 29466 документов только 2500 документов были переводами, то есть имели точные переводные эквиваленты на всех языках словаря.

В начале эксперимента был проведен анализ исходных текстов на предмет синонимии словарных единиц. Была проанализирована выборка из 847 английских, 313 французских и 289 немецких заголовков, в которых содержатся переводы словосочетания «Официальный представитель МИД Российской Федерации». Следует отметить, что качество перевода находится на высоком

уровне, так как работа выполнялась квалифицированными переводчиками. В таблице 1 приведены данные о количестве употреблений (КУ) в текстах на английском, немецком и французском языках переводов словосочетания «Официальный представитель МИД Российской Федерации».

Таблица 1

Сведения о синонимии в параллельных текстах

Английский перевод	КУ	Немецкий перевод	КУ	Французский перевод	КУ
1	2	3	3	5	6
Spokesman of Russia′s Ministry of Foreign Affairs	452	Sprecher des Aussenministeriums Russlands	47	porte-parole [officiel] du MAE de la Russie	141
official Spokesman of Russia′s Ministry of Foreign Affairs	290	offizieller Vertreter des Aussenministeriums Russlands	41	porte-parole [officiel] du Ministère Russe des affaires étrangères	104
Russian Foreign Ministry [official] Spokesman	68	Amtssprechers des russischen Aussenministeriums	37	porte-parole du Ministère des affaires étrangères de la Fédération de Russie	30
[official] Russian Foreign Ministry Spokesman	20	offizieller Sprecher des Aussenministeriums Russlands	20	porte-parole officiel du MAE Russe	15

Продолжение таблицы 1

1	2	3	3	5	6
Official Spokesman of the Russian Foreign Ministry	8	Amtssprecher des Aussenministeriums Russlands	18	porte-parole du Ministère des affaires étrangères de la Russie	15
Official Spokesman for Russia′s Foreign Ministry	2	Vertreter des Aussenministeriums der RF	10		
RF Foreign Ministry′s Spokesman	1	offizieller Vertreter des russischen Aussenministeriums	4		
		Aussenamtssprecher	2		
		amtlicher Sprecher des Aussenministeriums der RF	1		
		amtlicher Sprecher des russischen Aussenministeriums	1		

Из таблицы 1 видно, что количество синонимов и частота их появления в текстах для разных языков сильно отличается. Так, в английском варианте есть одна форма словосочетания, превалирущая над всеми остальными, во французском – две такие формы, а в немецком языке – вообще нет. Подобная тенденция сохраняется и для других словарных единиц. В связи с этим, при формировании частотных словарей в английские и французские словари попадали слова с частотой, не меньшей ⅔ от количества предложений, а для немецких словарей порог был снижен до ½.

Дадим краткое описание результатов эксперимента.

Этап1. Из текстов автоматически были вычленены заголовки, перенумерованы. Причем все четверки предложений, являющиеся переводами друг друга, получили один порядковый номер, помимо этого предложение снабжалось еще номером языка (1- русский, 2 –

английский, 3- немецкий, 4 – французский). В таблице 2 приводится пример из четырех предложений.

Таблица 2

Фрагмент исходных параллельных текстов

395#1#Ответ официального представителя МИД России А.В.Яковенко на вопросы СМИ в связи с заявлением Министра внутренних дел Франции Д. де Вильпена относительно изготовления террористами в Панкисском ущелье (Грузия) химического и бактериологического оружия#
395#2#Alexander Yakovenko, the Spokesman of Russia's Ministry of Foreign Affairs, answers a media question regarding French Interior Minister Dominique de Villepin Statement concerning chemical and bacteriological weapons being made by terrorists in Pankisi Gorge, Georgia#
395#3#Antwort des Amtsprechers des Aussenministeriums Russlands A.W.Jakowenko auf Fragen der Massenmedien im Zusammenhang mit Erklaerung des Innenministers Frankreichs D. de Villepin hinsichtlich der Produktion von Terroristen der chemischen und bakteriologischen Waffen in der Pankissky Schlucht (Georgien)#
395#4#Réponse d'A.V.Yakovenko, porte-parole du MAE de la Russie, aux questions des médias concernant la déclaration de D. de Villepin, Ministre de l'Iintérieur de la France, à propos de la fabrication d'armes bactériologiques et chimiques par les terroristes dans les gorges de Pankissi (Géorgie)#

Этап 2. Выделение словосочетаний из русских предложений и составление по ним частотного словаря. Получен словарь объемом около 400 словарных единиц с частотой f>=2 (в таблице 3 приведен фрагмент из него).
Этап 3. Составление частотных словарей словосочетаний по английским, немецким и французским текстам, соответствующих каждому русскому словосочетанию (промежуточные результаты, не сохраняются). Установление переводных соответствий, формирование многоязычного словаря. Фрагмент словаря приведен в таблице 4. Каждая словарная статья состоит из ПОС-а русского словосочетания и словосочетаний на всех языках, между которыми вставлен разделитель # (решетка). Синонимичные словосочетания разделяются точкой с запятой.

При выборе формата записи многоязычного словаря нужно решать проблемы кодировки. Так как далеко не все системы поддерживают UNICODE, принята кодировка Кириллица (Windows).

Таблица 3

Фрагмент частотного словаря слов и словосочетаний, составленного по текстам на русском языке

министр иностранных дел россии	158	департамент информации и печати мид россии	7
официальный представитель мид россии	140	газета время новостей	6
министр иностранных дел российской федерации	37	министр иностранных дел франции	6
заместитель министра иностранных дел россии	22	генеральный секретарь нато	5
постоянный представитель россии при оон	10	госсекретарь сша	5
государственный секретарь сша	8	министерство иностранных дел российской федерации	5
заместитель официального представителя мид россии	8	президент сша	5
генеральный секретарь оон	7	заседание совета безопасности оон	5

Таблица 4

Фрагменты многоязычного словаря слов и словосочетаний, полученного автоматическим путем

агентств интерфакс#агентство Интерфакс#Interfax News Agency#Presseagentur Interfax#Agence Interfax
агресси#агрессия#agression#Aggression#agression
адаптаци#адаптация#adaptation#Adaptierung#adaptation#
……
ближн восток#Ближний Восток#Middle East#Nahost#Proche-Orient
ближневосточн урегулировани#ближневосточное урегулирование#Middle East settlement#nahoestlichen Regelung#reglement Proche-Oriental
……
государств#государство#State#Staat#Etat
государственн программ культурн обмен# государственная программа культурного обмена#state program of a cultural exchange#staatliche Programm des kulturellen Austausches#programme d'Etat de l'échange culturel
государственн регистраци#государственная регистрация#State recording#staatliche Registrierung#l'enregistrement d'Etat
государственн регулировани тариф#государственное регулирование тарифов#State regulation of the fares#staatliche Regelung der Tarife#réglage d'Etat des tarifs
государственн секретар сша#Государственный секретарь США#US Secretary of State#staatssekretaer der USA#Secrétaire d'Etat des Etats-Unis
государственн фельдъегерск служб российск федераци#Государственная Фельдъегерская служба Российской Федерации#State Courier Service of the Russian Federation#russischen Staatlichen Kurierdienst#Service des courrriers d'Etat de la Fédération de Russie
……
федеральн космическ агентств#Федеральное космическое агентство#Federal Space Agency#Foederale Raumagentur#Agence fédérale spaciale
федеральн миграционн служб#Федеральная миграционная служба#Federal Migration Service#Foederale Migrationsdienst#Service fédéral des migrations
федеральн налогов служб#Федеральная налоговая служба#Federal Taxation Service#Foederale Steuerdienst#Service fédéral fiscal

Полученный автоматическим способом словарь нуждается в редактировании человеком, так как при анализе было выявлено ~ 15% случаев неправильного установления переводных соответствий. Наряду с орфографическими ошибками в текстах (более всего в немецких) причиной является наличие во всех языках синонимичных словарных единиц или грамматических конструкций для выражения одного и того же понятия.

Чтобы получить приемлемое покрытие текстов запросов, словарь, полученный автоматическим путем, был пополнен за счет других словарей [2, 5, 6, 7, 10]. Объем словаря составил 20 тыс. словарных единиц.

В процессе эксперимента выяснилось, что переводные эквиваленты на разных языках расположены, как правило, контактно. Исключения из этого правила практически не влияют на результат.

3.3. Анализ запросов пользователей

Для выполнения второй части исследования – перевода запросов пользователей Интернет на английский, немецкий и французский языки, компанией Яндекс были предоставлены исходные данные: набор данных «Протоколы работы поисковой системы». Этот набор данных содержит выборку из реальных данных за одну регулярную неделю работы Яндекса. Из выборки было извлечено 175808 текстов запросов. Все запросы считались равнозначными и не соотносились с конкретными сеансами или пользователями.

С точки зрения постановки задачи интерес представляют запросы, которые ориентированы на поиск информации на иностранных языках. В поисковых системах с функцией кросс-языкового поиска должен быть специальный интерфейс, для того чтобы пользователь мог выбрать языки, на которые он желает перевести свой запрос. Конечно, хорошо бы предоставить также возможность получения перевода результата поиска на родном языке. Однако сейчас об этом говорить рано, так как ни одна из существующих систем машинного перевода не может в автоматическом режиме качественно перевести текст произвольной тематики.

Поскольку истинные намерения пользователей нам были не известны, то мы считали, что на поиск иноязычной информации ориентированы запросы, в которых наряду с русскими словами есть хотя бы одно слово на других языках. Таких запросов оказалось 17350, а после исключения дублей их осталось 17195. На один

запрос в среднем приходится 3,8 слова. Среди этих словосочетаний было ~ 10% ошибочных. Типичные ошибки следующие: неправильное написание слов из-за плохого знания русского и других языков (в основном, используется английский), пропуск нужных или вставка лишних букв, замена или перестановка букв, использование в русских словах латиницы, а в иноязычных кириллицы, отсутствие пробелов между словами или появление ненужных пробелов внутри слов, замена пробела на символ «_» (нижнее подчеркивание).

Многие пользователи в качестве запросов используют вопросительные предложения или включают малоинформативную (иногда даже нецензурную) лексику и слова, относящиеся к служебным частям речи (союзы, предлоги, частицы). Такая лексика не должна переводиться.

Была выявлена наиболее частотная лексика, содержащаяся в запросах пользователей, и проверено вхождение ее в многоязычный словарь. В таблице 5 приводятся самые частые русские слова, встречающиеся в запросах (исключены частые предлоги и союзы). Среди частых слов встречаются ошибки. Так, например, слово «руссификатор» встретилось 80 раз, а правильное слово «русификатор» - 75 раз.

Таблица 5

Фрагмент частотного словаря слов, составленного по запросам пользователей

1801 скачать	157 москва
425 программа	156 купить
374 драйвер	151 описание
362 игра	144 русский
353 бесплатный	140 фото
293 сайт	135 прошивка
259 мелодии	122 прохождение
239 код	118 инструкция
235 телефон	118 патч
185 песня	115 карта
160 цена	111 файл

3.4. Формализация и перевод запросов пользователей

Формализация и перевод русскоязычных запросов проводились в следующей последовательности:

1). Морфологический анализ тестов запросов для определения грамматической информации входящих в них слов.

2). Выделение из текстов запросов наименований понятий, выраженных словами и словосочетаниями.

3). Поиск в многоязычном словаре найденных словосочетаний.

4). Формирование запроса на иностранном языке: объединение непереводимой части запроса (иноязычные слова) и переводов слов и словосочетаний.

В большинстве поисковых систем допускается задавать запрос в виде набора ключевых слов и словосочетаний, разделенных пробелами, т. е. если не указываются явно логические операторы, то по умолчанию используется оператор AND и находятся только документы, содержащие все слова запроса. В некоторых системах по умолчанию используется оператор «нечеткое» AND, который задает условия поиска как нечто среднее между поиском по пересечению и объединению терминов.

Во многих поисковых системах можно применять скобки для формирования сложных запросов. Группа слов, заключенная в скобки, может рассматриваться как отдельный запрос, таким образом можно учесть словосочетания при поиске. Правда, в таких известных зарубежных системах как Google и Yahoo, наличие скобок в запросе не влияет на результаты выдачи документов. Использование синтаксических операторов нецелесообразно, так как в разных поисковых системах синтаксис языка запросов различен. Поэтому все знаки и операторы (+,-, |, & и т.д.) должны удаляться перед переводом запроса.

Таким образом, формализация запроса сводится к автоматическому выделению из его текста ключевых слов и словосочетаний, построении поисковых образов словосочетаний, поиска словосочетаний в многоязычном словаре и объединении переводов словосочетаний с частью запроса, сформулированного на других языках.

При поиске словарные единицы считались тождественными при условии совпадения их ПОС-ов. Слова и словосочетания, не найденные в словаре, в перевод запроса не включались. Конечно, можно было бы такие слова включать в запрос в транслитерированном виде, чтобы распознавать в иноязычных текстах, например, фамилии и имена, отсутствующие в словаре. Правда, для разных языков действуют свои правила транслитерации (ср. Yakovenko– анг. и Jakowenko –нем., Putin - анг., нем., Poutine - фр.), поэтому не все фамилии будут найдены в немецких или французских текстах.

Исходный массив запросов без дублей был обработан с помощью разработанного авторами комплекса программ. Из 17195 словосочетаний было переведено 12890, что составляет 75%. Количество переведенных запросом можно увеличить до 95-98%, если значительно пополнить словарь.

Таблица 7

Примеры переводных соответствий запросов на русском, английском, немецком и французском языках

1) толковый словарь on-line#explanatory dictionary on-line#erklärende Wörterbuch on-line#dictionnaire raisonné on-line#
2) MMF с голосом#voice MMF#Stimme MMF#voix MMF#
3) flash-игры скачать бесплатно#games flash download free#Spiele flash downloads kostenlos #jeux flash downloads gratuit#
4) развитие Ford в Европе#development Europe Ford#Entwicklung Europa Ford#développement Europe Ford#
5) устройство для нанесения клея ENANO 350#device spread ENANO 350#Gerät Auftragen des Klebers ENANO 350#installation application de la colle ENANO 350#
6) автомобильный dvd проигрыватель#automobile record player dvd#Autoplattenspieler dvd#tourne-disques d'automobile dvd#
7)что такое маркер Exif в Jpeg#marker Exif Jpeg#Marker Exif Jpeg#marqueur Exif Jpeg#
8) цена на giga-byte 8 ip1000#price giga-byte 8 ip1000#Preis giga-byte ip1000#prix giga-byte 8 ip1000#

4. Выводы

1. Проведенное исследование показало, что автоматизированные методы создания словарей применимы в многоязычной лексикографии. Авторы исследования разработали методы составления многоязычных словарей и реализовали их в виде программного продукта, на вход которого поступают параллельные тексты заголовков на четырех языках (русском, английском, французском, немецком), а на выходе имеется многоязычный словарь слов и словосочетаний.

Метод был апробирован на текстах новостей дипломатии. Естественно, необходимо обрабатывать тексты по разным тематикам, причем больших объемов, а также формировать вспомогательные словари, необходимые при анализе текстовой информации. В наборе таких словарных средств обязательно

должны быть многоязычные словари персон (имена и фамилии), словари географических названий, словари названий организаций, словари сокращений. Некоторые из перечисленных словарей уже есть в книжной форме, некоторые готовятся к изданию как, например, многоязычный географический словарь. Мы использовали подобные словари, но вопросов их разработки не касались. Создать словарь персон вручную – задача нереальная. В качестве исходных текстов также можно использовать параллельные тексты. Вначале нужно провести транслитерацию с русского на английский язык, а потом найти в английских текстах слова, в точности совпадающие с транслитерированными русскими словами. Далее соответствие между русскими словами и транслитерациями на других языках можно устанавливать, исходя из написаний английских слов и применяя правила уже для найденных слов.

2. Отметим еще один аспект рассматриваемой проблемы – возможность составления словарей не только по заголовкам, но и по полным текстам документов. Однако для этого необходимо располагать процедурами морфологического анализа и словарными средствами для всех рабочих языков составляемого словаря. Работы такого масштаба под силу только большим коллективам специалистов, перед которыми стоит цель создания системы машинного перевода или организации кросс-языкового поиска информации.

3. При организации кросс-языкового поиска в Интернет, прежде всего, необходимо провести анализ пользовательских запросов. Многоязычный словарь должен включать в свой состав лексику, приближенную к реальным текстам и запросам. Анализ запросов показал, что пользователи используют часто нетрадиционную лексику, но которая, тем не менее, должна найти свое место в словаре, исключая нецензурную лексику. В запросах часто используются малоинформативные слова. При переводе пользовательских запросов не имеется в виду дословный перевод со всеми служебными словами и знаками препинания. Из запроса вычленяются слова и словосочетания, определяющие его смысл, и только они переводятся на другие языки.

4. Эксперимент по переводу запросов показал, что словарь в 20000 лексических единиц обеспечивает перевод ~ 75% запросов. Чтобы обеспечить достаточную полноту перевода пользовательских запросов, словарь должен быть пополнен наиболее часто используемой в Интернете лексикой: прежде всего из областей программного и аппаратного обеспечения, индустрии культуры и туризма.

5. Литература

1. Абрамова Н.Н., Бевзенко Е.А. Составление словарей словосочетаний по неформализованным текстам. - Вопросы информационной теории и практики, № 53, ВИНИТИ, 1985.
2. Абрамова Н.Н. и др. Многоязычный электронный словарь по внешней политике. - Материалы 6-ой международной конференций НТИ-2002 - Москва, 16-18 октября 2002 г.
3. Белоногов Г.Г. и др. Компьютерная лингвистика и перспективные информационные технологии- М.: Рус. мир, 2004. -248 с.
4. Беляева Л.Н. Лексикографический потенциал параллельного корпуса текстов. Международная конференция "Корпусная лингвистика- 2004". Санкт-Петербург, 12-14 октября 2004 г.
5. Борковский А.Б. и др. Словарь по программированию (английский, русский, немецкий, французский) - М.: Рус. яз.,1991. – 286 с.
6. Горский В.А. и др. Глоссарий по Европейской интеграции. Термины договоров соглашений Европейского Союза на английском, русском, французском, немецком и нидерландском языках.- М.: Интердиалект, 1998.- 357 с.
7. Милорадович Живан М. Словарь русско-английский, немецкий, французский, англо-, немецко-, французско-русский словарь.- М.: Вече, 1997.-560 с.
8. Поминов А.В. Некоторые вопросы построения многоязычных автоматических словарей/ Материалы Диалога 2001. Том 2. http://www.dialog-21.ru/archive_article.asp?param=7023&y=2001&vol=6078
9. Светова С.Ю. Опыт создания онлайнового переводчика ПРОМТ/ Материалы Диалога 2000. Том 2. http://www.dialog-21.ru/full_digest.asp?digest_id=18187
10. Словарь экономических терминов на 11 языках. – М.: изд-во «АСТ», 2004.-1344 с.
11. Levin R. Tools for Multilingual Communication.-Multilingual Computing&Technology, #70 Volume 16 Issue 2.

---

[1] Цепочка из словообразовательной основы опорного слова словосочетания (обычно первого слева существительного) и словообразовательных основ всех остальных слов, упорядоченных по алфавиту.

[2] Этот алгоритм и все словарные средства для построения процедур разработаны авторами с участием студента МГТУ им. Баумана Абрамова В.Е. с использованием подходов и идей проф. Г.Г. Белоногова и д.т.н. Ю.Г. Зеленкова [3].